I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

The Policy Infrastructure for Big Data: From Data to Knowledge to Action

FARNAM JAHANIAN*

I. INTRODUCTION

Innovative information technologies are transforming the fabric of society, and data is everywhere. They are produced in rapidly increasing volume and variety by virtually all scientific, educational, governmental, societal and commercial enterprises.

Advances in data collection, storage capacity, computational speeds, and analytical tools are yielding new methods of observation, discovery, access and analysis. These advances are enabling an increased capacity to extract information, reveal previously unknown correlations, generate hypotheses and infer meaning from data. This capacity is expected to grow, and is already disrupting the status quo across all sectors. [1]

This essay describes new opportunities presented by Big Data, including their evolving impact on research, discovery and innovation. It also outlines the policy infrastructure that is enabling all sectors to harness the power of Big Data, as well as some of the challenges that lie ahead.

---

* Farnam Jahanian, PhD, is the Vice President for Research at Carnegie Mellon University. This work was done while the author served as Assistant Director for Computer and Information Science and Engineering at the National Science Foundation.

[1] "Dealing with Data," *Science Magazine* 331, no. 6018 (2011): 692-693; "Economist Special Report: The Data Deluge," *The Economist*, February 25, 2010, http://www.economist.com/node/15579717.

II. WHAT IS BIG DATA?

The term "Big Data" refers not only to the enormous volume of data being generated from a range of sources, but also its heterogeneity, complexity and diversity, as well as the rate at which it is generated.[2] This resource, along with new paradigms and capabilities for data access, management and analysis, is changing the world.

We are living in the *era of data and information*, characterized by an explosive growth in the size, complexity, and velocity of digital data that is generated and collected. This era is enabled by modern experimental methods and observational studies; large-scale simulations; scientific instruments such as telescopes and particle accelerators; Internet transactions, email, videos, images, and click streams; and the widespread deployment of sensors everywhere in the environment, in our critical infrastructure such as bridges and smart grids, in our homes, and even on our clothing. Consider what happens in just one minute on the Internet. In 2013, the average minute saw over 240,000 photos uploaded to Facebook, 350,000 tweets, 100 hours' of video uploaded to YouTube and 3.5 million Google searches.[3]

Such activity is increasingly being conducted via mobile devices, which are becoming the most accessible digital systems on the planet. By the end of 2014, mobile devices are expected be the world's predominant means for accessing the Internet, surpassing traditional computers as access points. The number of mobile devices alone will soon exceed the number of people on earth, currently more than 7 billion. This growth is fueled by the emergence of new apps for mobile devices such as smart phones, as well as the rise of social media.[4] The amount of data moving through mobile networks world-wide is expected to rise at a compound annual growth rate (CAGR) of 61% to

---

[2] The National Science Foundation. *Critical Techniques and Technologies for Advancing Big Data Science and Engineering (BIGDATA),* http://www.nsf.gov/pubs/2014/nsf14543/ nsf14543.htm; Victor Mayer-Schonberger and Kevin Cukier. *Big Data: A Revolution That Will Transform how We Live, Work and Think* (New York: Houghton Mifflin Harcourt, 2013), 2.

[3] Leo Mirani, "A Snapshot of One Minute on the Internet, Today and in 2012," *Quartz*, November 26, 2013, http://qz.com/150861/a-snapshot-of-one-minute-on-the-internet-today-and-in-2012/#150861/a-snapshot-of-one-minute-on-the-internet-today-and-in-2012/.

[4] Christopher Surdak, *Data Crush: How the Information Tidal Wave Is Driving New Business Opportunities* (New York: AMACOM, 2014), 9-22.

15.9 exabytes per month by 2018, with roughly two-thirds due to video content,[5] fueling the digital data wave.

Much of today's vast quantities of digital data is unstructured or inexact. According to the recent book, *Big Data: A Revolution That Will Transform how We Live, Work and Think,* we have historically relied on "information that is small, exact, and causal in nature…," but our capabilities are changing now that "data is huge, can be processed quickly, and tolerates inexactitude."[6]

This has three significant implications for how value may be extracted from data. First, the ability to collect and manage large data sets means that researchers and analysts are not limited to sampling; all of the data may now be at play. Second, every piece may not be accurate or precise on its own, but when combined, broad insights may nonetheless be gained. Finally, the availability of myriad data sets enable the identification of patterns that indicate useful correlations but not causality; these correlations may be used to generate new hypotheses about the interrelation of variables.[7]

III. WHY BIG DATA MATTERS

Big Data is important to all facets of the discovery and innovation ecosystem, including the nation's academic, government, industrial, entrepreneurial, and investment communities.

First, Big Data has profound implications for the economy. Insights and more accurate predictions from large and complex data sets drive creation of new products and services, boost the productivity of businesses, and potentially transform business models. The data technology and services market alone is expected to grow from $3.2 billion in 2010 to $32.4 billion in 2017.[8] Data in itself is

---

[5] "Global Mobile Data Traffic Forecast Update, 2013-2018," *Cisco*, February 5, 2014, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[6] Viktor Mayer-Schonberger and Kenneth Cukier, Big Data: A Revolution That Will Transform how We Live, Work and Think (New York: Eamon Dolan/Mariner Books, 2013), 16.

[7] Ibid., 15.

[8] "IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy," *Business Wire*, March 7, 2012, http://www.businesswire.com/news/home/20120307005036/en/IDC-Releases-Worldwide-Big-Data-Technology-Services; "Worldwide Big Data Technology and Services 2013-2017 Forecast," *IDC*, accessed May 27, 2014, http://www.idc.com/getdoc.jsp?containerId=244979.

now being monetized and sold to third parties due to its value in profiling customers' preferences for use in predictive analysis and targeted advertising. Amazon, eBay and Groupon, and even many brick-and-mortar retailers such as Target, all employ predictive data analysis to enhance services and better meet customer needs.[9]

Second, advances in our ability to store, integrate, and analyze data are accelerating the pace of discovery in almost every science and engineering discipline. For example, the amount of data collected by the Sloan Digital Sky Survey (SDSS)[10] in its first few weeks of operation in 2000 was greater than it had been gathered in the entire history of astronomy.[11] Within a decade, over 140 terabytes of information was collected,[12] representing 35% of the sky.[13] The planned Large Synoptic Survey Telescope (LSST), with construction anticipated to be complete in 2022, will eclipse this effort in its first week of operation. This project is expected to generate 40 terabytes of data each night and cover half of the sky every three days, providing vast new opportunities to study dark matter and energy, map near-Earth asteroids and identify transient events such as novae or supernovae.[14]

In addition to large instrument-driven science, the "long tail" data collected by hundreds of thousands of scientists through experiments, simulations, sensors and surveys collectively represent an enormous but largely untapped scientific resource. Combining the data from researchers working in small or disconnected groups will enable new and collaborative discoveries.

Finally, data analytics have the potential to solve some of the Nation's most pressing challenges, yielding enormous societal benefit in priority areas such as health care and well-being, sustainability,

---

[9] Christopher Surdak, *Data Crush: How the Information Tidal Wave Is Driving New Business Opportunities* (New York: AMACOM, 2014), 74, 76.

[10] "The Sloan Digital Sky Survey: Mapping the Universe," *Sloan Digital Sky Survey*, http://www.sdss.org/.

[11] Viktor Mayer-Schonberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform how We Live, Work and Think,* (New York: Eamon Dolan/Mariner Books, 2013).

[12] "SDSS Data Release 4," *Sloan Digital Sky Survey*, last modified 2005, http://classic.sdss.org/dr4/index.html.

[13] "SDSS," *Universe* News, http://www.universenews.in/sloan-digital-sky-survey/.

[14] Michael Strauss, "LSST Science Book, Version 2.0," *Cornell University Library*, December 1, 2009, http://arxiv.org/abs/0912.0201; "Large Synoptic Survey Telescope," http://www.lsst.org/lsst/.

energy, transportation, education and cybersecurity. From new knowledge about protein structure paving the way for advances in biomedical research and clinical decision-making, to new ways to mitigate and respond to natural disasters, to new strategies for effective learning and education, there are enormous opportunities for data to improve quality of life and lay the foundations for American competitiveness for decades to come.

Through long-term, sustained investments in foundational computing, communications and information technology research, and the development and deployment of large-scale facilities and cyber infrastructure, Federal agency R&D investments over the past several decades have both helped generate the explosion of data as well as advance our ability to capture, store, analyze and use these data for societal benefit. More specifically, we have seen fundamental advances in machine learning, knowledge representation, natural language processing, information retrieval and integration, network analytics, computer vision, and data visualization, which together have enabled Big Data applications and systems that have the potential to transform all aspects of our lives.

These investments are already starting to pay off, demonstrating the power of Big Data approaches across science, engineering, medicine, commerce, education, and national security.

For example, researchers have used data analytics to pioneer a self-programming thermostat that uses activity sensing and machine learning to automate the most use-efficient, cost-cutting and energy efficient home heating and cooling routines. Such systems create savings for consumers[15] and reduce home energy consumption (responsible for more than 20% of the Nation's energy usage[16]), nearly half of which is used for heating and cooling.[17]

In the transportation sector, a number of regional ventures – in Los Angeles, the Bay Area, northern New Jersey, and in Washington, DC region – are integrating heterogeneous data sources such as road sensors, traffic cameras, individuals' GPS devices, etc., to develop principles and methods that go beyond real-time traffic data and

---

[15] "Saving Energy," *Nest*, http://www.nest.com/saving-energy/.

[16] "Annual Energy Outlook 2014 – Market Trends: U.S. Energy Demand," U.S. Energy Information Administration (EIA), http://www.eia.gov/forecasts/aeo/ MT_ energydemand.cfm# indus_comm.

[17] "Heat and Cool Efficiently," *Energy Star*, http://www.energystar.gov/ index.cfm?c=heat_cool.pr_hvac.

enable the inference of traffic patterns over entire cities.[18][19][20] In Los Angeles, for example, city planners have synchronized every one of the 4,500 traffic signals across 469 square miles of downtown. They use a system of sensors, live traffic cameras, and a centralized computing platform to make constant, automated adjustments and to keep cars running as smoothly as possible.[21] Under this system, the average speed of traffic across the city has increased by 16%, with delays at major intersections down 12%.

Data analytics are also already improving healthcare by aiding in the diagnosis of diseases such as breast cancer, the second-most common cancer among American women.[22] Out of over 6,000 possible features, researchers who applied image analysis techniques to hundreds of breast cancer biopsy images were able to identify a small subset of cellular features that were predictive of survival time among breast cancer patients. Unexpectedly, the features that were the best predictors of patient survival were not from the cancer tissue itself, but rather from adjacent tissue, a correlation that had gone undetected by pathologists and clinicians. [23] These new discoveries will allow clinicians to better understand the genesis and morphology of breast cancer, enabling personalized treatments that aim to improve survival times among patients.

---

[18] Ari Entin, "IBM, Caltrans and UC Berkeley Aim to Help Commuters Avoid Congested Roadways Before their Trip Begins," *IBM,* April 13, 2011, http://www-03.ibm.com/ press/us/en/pressrelease/34261.wss.

[19] Shira Ovide, "Trapping 'Big Data' to Fill Potholes: Start-Ups Help States and Municipalities Track Effects of Car Speeds, Other Variables on Traffic," *The Wall Street Journal*, June 12, 2012, B6.

[20] "RITIS," CATT Lab: A User-Focused R&D Laboratory at the University of Maryland, http://www.cattlab.umd.edu/?portfolio=ritis.

[21] Ian Lovett, "To Fight Gridlock, Los Angeles Synchronizes Every Red Light," *The New York Times*, April 2, 2013, A11.

[22] "What are the Key Statistics about Breast Cancer?*" American Cancer Society,* last modified January 31, 2014, http://www.cancer.org/cancer/ breastcancer/detailedguide/ breast-cancer-key-statistics.

[23] Andrew H. Beck, etl al., "Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival," *Science Transnational Medicine* 3, no. 108 (2011): 108.

IV. A NEW ERA OF SCIENTIFIC DISCOVERY AND ENGINEERING INNOVATION

These kinds of breakthroughs are catalyzing a profound transformation in the culture and conduct of scientific research, requiring new methods to derive knowledge from the data and new infrastructure to manage, curate and serve data to communities. Big Data will also necessitate new approaches to education and training and new types of collaborations between multi-disciplinary teams and communities that have the potential to solve today's most complex science and engineering challenges.

Data-driven discovery is revolutionizing scientific exploration and engineering innovations. This approach has been called the "fourth paradigm," in contrast to the three earlier modalities of scientific research: empirical observation and experimentation, analytical/theoretical approaches, and computational science and simulation.[24] The data-driven approach not only complements these earlier approaches, but has the promise to revolutionize science even further.

Indeed, the fourth paradigm has led to improved hypotheses and faster insights. While data access and analysis are already having enormous impacts, the opportunities for the future are immense. Imagine a day when:

- By integrating biomedical, clinical, and scientific data, we can predict the *onset of diseases and identify unwanted drug interactions*.

- By *accurately predicting natural disasters* such as hurricanes and tornadoes, we can employ life-saving and preventative measures that mitigate their potential impact.

- By integrating emerging technologies, such as Massively Open Online Courses (MOOCs) and inverted classrooms, with knowledge from research about how people learn, we can *transform formal and informal education*.

---

24 Tony Hey, Steward Tansley, and Kristin Tolle, eds. "The Fourth Paradigm: Data-Intensive Scientific Discovery," (Microsoft Corp. 2nd ed. 2009), available at http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx.

- By *correlating disparate data streams* through text mining, image analysis, and face recognition, we can enhance public safety and security.

Many research and development challenges remain. Below is a list of hard problems in Big Data that the research community is addressing:

- *Many data sets are too poorly organized to be usable*. Research must come up with new techniques to better organize and retrieve data.

- *Many data sets are comprised of unstructured or incomplete data*. Research must develop new data mining tools and/or machine learning techniques to make these data usable. Opening government data to all is one of the first steps to spur innovation.

- *Many data sets are heterogeneous in type, structure, semantics, organization, granularity, and accessibility*. Research must find novel ways to integrate and customize access to federated data, and to make heterogeneous data more interoperable and usable.

- *The utility of data is limited by our ability to interpret and use it*. Research must find better usability techniques to extract and visualize actionable information. Research needs to discover new techniques for evaluating and showing results.

- *More data are being collected than we can store*. With the right data infrastructure, practitioners could analyze data as it becomes available; they could immediately decide what to archive and what to discard.

- *Many data sets are too large to download or send over today's Internet*. With the right data infrastructure, practitioners could analyze the

data wherever it resides, instead of sending it to data centers.

- *Distributed storage and cloud computing pose new challenges to the integrity and security of data sets.* New strategies or protocols could ensure security and reliability of datasets, regardless of where they are housed or analyzed.

- *Large and linked data sets may be exploited to identify individuals.* Research on privacy protection and Big Data is critical; new techniques and analysis could have "built-in" privacy preserving characteristics.

The landscape of open research and development challenges is vast, and tackling them will be critical to the ability of all sectors to seize the opportunities afforded by this new, data-driven revolution. The government has long played a critical role in seeding efforts to solve such confounding problems, leading directly to transformative new technologies with great economic and societal benefit.[25] Enabling these discoveries and innovations has been and will continue to be a national priority.

V. THE U.S. NATIONAL BIG DATA RESEARCH AND DEVELOPMENT INITIATIVE

In December 2010, the President's Council of Advisors on Science and Technology (PCAST) published a report to the President and Congress entitled *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*.[26] In the report, PCAST pointed to the research challenges involved in large-scale data management and analysis and the critical role of Networking and Information Technology (NIT) in moving from data to knowledge to action, underpinning the nation's future

[25] Mazzucato, Mariana, "The Entrepreneurial State," *Soundings*, no. 49 (2011): 131-142.

26 "Designing a Digital Future: Federal Funded Research and Development in Networking and Information Technology" Report to the President and Congress, December 2010, http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf.

prosperity, health and security. The report recommended several actions to take advantage of Big Data opportunities.

The Office of Science and Technology Policy (OSTP) in the Executive Office of the President (EOP) responded to these recommendations, in part, by chartering a Big Data Senior Steering Group (BDSSG) that would focus on Big Data R&D under the umbrella of the Networking and Information Technology R&D (NITRD) program. The BDSSG coordinates Big Data R&D across the member agencies by 1) promoting new science and accelerating the progress of discovery through large, heterogeneous data; 2) exploiting the value of Big Data to address areas of national need, agency missions and societal and economic importance; 3) supporting responsible stewardship and sustainability of Big Data resulting from federally-funded research; and 4) developing and sustaining the infrastructure needed to advance data science.[27]

Currently, the National Science Foundation (NSF) and the National Institutes of Health (NIH) co-chair the BDSSG. Membership is comprised of representatives from the science agencies, including the Department of Energy's (DOE) Office of Science, the Department of Homeland Security (DHS), the National Aeronautics and Space Administration (NASA), NIH, the National Institute of Standards and Technology (NIST), NSF and the U.S. Geological Survey (USGS).
After its establishment, the BDSSG inventoried existing Big Data programs and projects across the agencies and began coordinating their efforts in four main areas: investments in Big Data core techniques and technologies, education and workforce, domain cyber infrastructure, and challenges and competitions.[28] Other critical areas for Big Data were also identified, including privacy issues, open access to government data, and partnerships with industry and not-for-profits.

On March 29, 2012, the Administration launched the National Big Data Research & Development Initiative, with six Federal departments and agencies announcing more than $200 million in new commitments for Big Data R&D.[29] Now in its third year, this OSTP-led Initiative aims to advance the tools and techniques used for Big Data analysis and the human capital needed to move from data to knowledge to action

---

[27] "Big Data Senior Steering Group (BDSSG)," *The Networking and Information Technology Research and Development (NITRD) Program,* http://www.nitrd.gov/nitrdgroups/index. php?title=Big_Data_%28BD_SSG%29#title.

[28] Ibid.

[29] Tom Kalil, "Big Data is a Big Deal," *The White House*, March 29 2012, http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

In its second year, the BDSSG encouraged multiple stakeholders, including private industry, academia, state and local government, non-profits and foundations, to develop and participate in Big Data innovation projects across the country, emphasizing the value and importance of building multi-stakeholder partnerships in all areas of Big Data science and engineering. On November 12, 2013, a NITRD/OSTP event[30] highlighted 30 new, innovative collaborations involving 90 partners[31] that are enabling the transition from data to knowledge to action, all with great potential benefit for the Nation.

The National Big Data Research and Development Initiative has already spurred an impressive range of new programs and projects throughout Federal agencies.[32] This committed support continues. For example, The National Institutes Health recently launched the multiyear Big Data to Knowledge (BD2K) initiative,[33] to develop new methods, tools and capacities for enabling biomedical scientists to harness the potential of the big data currently being generated by the research community, with the potential to transform our knowledge and strategies for diagnosing, treating and curing disease. DARPA's XDATA program is investing in the development of computational techniques and software tools for analyzing imperfect, incomplete or unstructured data. DARPA has also created an "Open Catalogue" of open source software and research publications generated through the program.[34] A final example is NSF's BIGDATA solicitation, "Critical Techniques and Technologies for Advancing Big Data Science &

---

[30] "'Data to Knowledge to Action' Event Highlights Innovative Collaborations to Benefit Americans," *Press Release, Office of Science and Technology Policy*, November 12, 2013, available at http://www.whitehouse.gov/ sites/default/files/microsites/ ostp/Data2Action%20Press%20Release.pdf.

[31] "Fact Sheet, Data to Knowledge to Action: New Announcements," November 12, 2013, http://www.whitehouse.gov/sites/default/files/microsites/ostp/Data2Action%20Announcements.pdf.

[32] See e.g., "Fact Sheet, Data to Knowledge to Action: Progress by Federal Agencies," November 12, 2013, http://www.whitehouse.gov /sites/default/files/microsites/ ostp/Data2Action%20Agency%20Progress.pdf.

[33] "NIH Big Data to Knowledge," National Institutes of Health, http://bd2k.nih.gov/about_bd2k.html#sthash.LEXknlor.dpbs.

[34] See, "Big Data: Seizing Opportunities, Preserving Values," Executive Office of the President, May 2014, http://www.whitehouse.gov/ sites/default/files/docs/ big_data_privacy_report_5.1.14_final_print.pdf; "Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million in New R&D Investments," Press Release, Office of Science and Technology Policy, March 29, 2012, http://www.whitehouse.gov/ sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

Engineering," which supports development of both novel foundational techniques and new innovative applications of techniques and technologies to advance our capacity for data-driven discovery.[35]

V. A FRAMEWORK FOR NATIONAL SCIENCE FOUNDATION INVESTMENTS

NSF expects that improvements in access, manipulation, data mining, management, analysis, sharing and storing of Big Data will provide new insights, change paradigms of research and education, and create new approaches to addressing national priorities. NSF has identified four major investment areas that address these challenges and serve as the foundations of a comprehensive, long-term agenda. They are:

A. *Foundational Research in all Areas of Science and Engineering:*

Advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets. Facilitate the development of new data analytic tools and algorithms; scalable, accessible, and sustainable data infrastructure; and large-scale integrated statistical modeling. Advance our knowledge and understanding of mathematical and physical systems, the science of learning, and human and social processes and interactions.

B. *Cyberinfrastructure*:

Provide science, engineering, and education with a comprehensive data infrastructure that will enable the capture, management, curation, analysis, interpretation, archiving and sharing of data of unprecedented scale, parallelism, and complexity in a manner that will stimulate discovery in all areas of inquiry, and from all instruments and facilities, ranging from campus- to national-level investments.

C. *Education and Workforce Development*:

Ensure that the future, diverse workforce of scientists, engineers, and educators is equipped with the skills to make use of, and build

---

[35] See, "Critical Techniques and Technology for Advancing Big Data Science and Engineering," National Science Foundation, http://www.nsf.gov/funding/pgm_summ.jsp? pims_id=504767.

upon, the next generation of data analytics, modeling, and cyber infrastructure. Support new approaches to K-16 teaching and learning that takes advantage of new computation- and data-driven approaches.

D. *Scientific Community Building and Governance:*

Support transformative interdisciplinary and collaborative research in areas of inquiry stimulated by data through the development of robust, shared resources and partnerships across diverse communities. Acknowledge the new challenges surrounding reproducibility, storage, curation, and open dissemination of scientific data in all its forms, and recognize its importance for accelerating fundamental discovery, interdisciplinary research, and innovation in society. Open and shared data can enable new approaches for communities to address complex problems in science and engineering.

NSF is taking a bold and comprehensive approach for this new data-centric world; from fundamental mathematical, statistical and computational approaches needed to understand the data, to infrastructure at a national and international level needed to support and serve our communities, to policy enabling rapid dissemination and sharing of knowledge. Together, these activities will accelerate scientific progress, create new possibilities for education, enhance innovation in society and be a driver for job creation.

VI. CHALLENGES AHEAD

As the major science and engineering challenges are tackled, it is critical that we address several key social and policy dimensions of the Big Data revolution.

First, new technological capabilities and the increasing amounts of data generated will challenge existing models for accessing data, an issue intertwined with policy. In general, researchers want access to data to fuel their discoveries. Public policy requires access to data, but also the protection of privacy, intellectual property, and other sensitive information. These considerations also affect the roles of publishers and scientific societies.

On February 22, 2013, a memo from OSTP directed U.S. Federal Agencies with more than $100 Million in annual R&D expenditures to develop a plan to support "increased public access" to the results of

their research.[36] This includes both peer-reviewed publications and digitally formatted scientific data resulting from unclassified research. Agency plans were submitted in the fall of 2013, and efforts are underway to develop new policies and resources to make more research data available.

Second, the increasing quantities of data gathered from myriad sources raise significant questions about the preservation of personal privacy. Privacy is a complex issue that lies at the intersection of personal and cultural values, expression, communication, technology, and law. New technologies are enabling the capture of more and more potentially sensitive data. They are also enabling extraction or inference of information that was not previously recoverable from a given data set.

While shopping, browsing, or using social media on the Internet, our preferences, habits and interactions are commonly tracked and stored. Customers can even negotiate away their privacy by sharing personally identifiable information in exchange for retailer discounts. "Terms of Use" for online services often formally condition use of a tool or service on an individual's agreement to allow use of his or her personal information. Surveillance and traffic cameras collect images, often without the subject's knowledge or consent. Wireless service providers track customers' phone connections and the locations from which they are made. Even seemingly benign data, when coupled with secondary data sources, have the potential to reveal personal information.[37]

Nonetheless, these same analytical tools promise enormous societal benefit. Not only that, but technology may offer effective solutions to new or existing challenges to personal privacy; future technologies could be designed with rigorous privacy protections embedded. The role of technology in privacy preservation should be an active area of research in the field of Big Data analytics.

On January 17, 2014, the President announced a 90-day White House review of Big Data and privacy. The review resulted in two important reports. The first, produced by PCAST, *Big Data: A Technological Perspective,*[38] explores the relationship between Big

---

[36]John Holdren, "Memorandum for the Heads of Executive Departments and Agencies," Executive Office of the President Office of Science and Technology, Released: February 22, 2014, http://www.whitehouse.gov/ sites/default/files/microsites/ ostp/ ostp_public_access_memo_2013.pdf.

[37] Steve Lohr, "How Privacy Vanishes Online" *The New York Times*, March 16, 2010, http://www.nytimes.com/2010/03/17/technology/17privacy.html?_r=0.

[38] John Holdren, Graham, Susan L., Press, William, "PCAST Releases Report on Big Data and Privacy," Office of Science and Technology Policy at The White House, May 1, 2014,

Data technologies and privacy preservation. The second, *Big Data: Seizing Opportunities, Preserving Values*,[39] addresses how Big Data affects daily life and the changing roles of citizens and their governments and how best to maximize its potential at minimal risk to privacy. A recent book, *Privacy, Big Data and the Public Good*,[40] proposes conceptual, practical and statistical frameworks for understanding and addressing the interplay between Big Data and privacy.

Finally, the nation faces the challenge of securing a workforce equipped to lead in the data age. A report by the McKinsey Global Institute estimated, "[b]y 2018 the United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data."[41] Preparing and sustaining a strong workforce will require sustained investments in Science, Technology, Engineering and Mathematics (STEM) Education, creation of solid tracks for professionalization in data analytics and information technology, and inspiration for the next generation of innovators.

As mentioned above, education and workforce development is a key pillar of NSF's investment framework for Big Data. Improving the Nation's capacity in data science is also an Agency Priority Goal.[42] NSF's education and training activities include the Advanced Technological Education program,[43] the Innovative Technology Experiences for Teachers and Students program,[44] the Research

---

http://www.whitehouse.gov/blog/2014/05/01/pcast-releases-report-big-data-and-privacy.

[39] "Big Data: Seizing Opportunities, Preserving Values," Office of Science and Technology Police at The White House May 2014, http://www.whitehouse.gov /sites/default/files /docs/ big_data_privacy_report_5.1.14_final_print.pdf.

[40] Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, *Privacy, Big Data and the Public Good* (Cambridge University Press: 2014).

[41] James Manyika, et al "Big data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey & Company, May 2011, http://www.mckinsey.com/ insights/business_technology/big_data_the_next_frontier_for_innovation.

[42] National Science Foundation, *Strategic Plan for 2014-2018,* March 2014 http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf14043.

[43] National Science Foundation. *Advanced Technological Education Program,* http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5464.

[44] National Science Foundation. *The Innovative Technology Experiences or Students and Teachers Program*, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5467.

Experience for Undergraduates program,[45] the NSF Research Traineeship (with an emphasis on Data-Enabled Science and Engineering),[46] and CAREER Awards[47] to support junior faculty who are exemplary teacher-scholars, both performing research and implementing excellent education plans for their students.

These issues are keystones of the Big Data research and development landscape. In order to sustain technological advances of benefit to society, solutions addressing data access, privacy and workforce development must be included in strategies throughout public, private and academic enterprises.

VII. LOOKING FORWARD

The rich and abundant data currently enabled by today's technologies provide a transformative new currency for science, engineering, business and government. Just as important, pervasive access to mobile devices, broadband connections, social media and other advanced cyber infrastructure comprise a rich ecosystem for communication and collaboration. Citizens, scientists and educators alike now communicate by sharing data; not only raw data, but also emails, software, publications, reports, simulations and visualizations. Coupled with appropriate policy and infrastructure development, this can create a new and profound ability to combine the efforts and resources of researchers at multiple scales to address far more complex grand challenge problems of science and society than was possible before.

Decades of technological advances and emerging social and policy interests have converged to open up a new landscape of opportunity. Moving forward, the government must play a continued role in investing in foundational research, seeding new knowledge and technological capabilities that will fuel innovation and growth across all sectors.

With robust, sustained public and private support, data-driven discovery and decision making will help position the Nation at the forefront of advances in science and engineering, enhancing economic prosperity, national security, and quality of life for future generations.

---

[45] National Science Foundation. *The Research Experience for Undergraduates Program*, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5517&org=NSF.

[46] National Science Foundation. *NSF Research Traineeship,* http://www.nsf.gov/pubs/2014/nsf14548/nsf14548.htm.

[47] National Science Foundation. *CAREER Awards,* http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503214.